Analysis of traffic tickets for 2019 in Bogotá Colombia

Julián Leonardo Martínez Camargo

Mayo 14, 2021

1. Introduction

For the final Project of the Coursera IBM Data Science Course, first it was necessary to define the question to establish what type of data is required to describe the situation accurately, then implementing statistics to measure the behavior and impact of the data. Finally, to generate a conclusion that explains the analysis of the results.

1.1.Business Problem

Bogota is a very crowded city and the number of vehicles per person is growing considerably, it is important to control this increasing mass of vehicles to reduce the number of accidents and incidents related to them. Tickets are an accurate variable to measure which are the most common infractions in the city. I have a particular interest in knowing if the type of bar, pub or other categories of establishments have some influence on the number of tickets related to drivers under the effects of alcohol. This information could be useful for the authorities to design plans where the checkpoints are better-located guaranteeing a safer city.

2. Data sources

2.1.Dataset download

In the following link you can download the dataset corresponding to the comparisons made in the city of Bogotá in 2019

https://datosabiertos.bogota.gov.co/dataset/comparendos-2019-bogota-d-c

2.2. Data definition

The dataset contains 402.500 records, in these data we find the following fields

- Longitud (Longitude): Spot meridian
- Latitud (Latitude): Spot parallel
- Objectid (Id row): primary key of the dataset
- Placa (Plate): vehicle identification
- Num_comparendo (traffic ticket): traffic ticket identifier
- Fecha (Date time): Date of the event
- Hora (Time occurrence): Time of the event
- Mes (Month): Month of the event
- Medio_deteccion (detection): How was the traffic ticket taken
- Clase_vehiculo (Vehicle Class): motorcycle, car, truck or other
- Tipo_servicio (Type Service): public, private or other
- Infraccion (Infraction):Type of Traffic ticket
- Valor (Value): Value to pay for the ticket
- Localidad (Borough): City area
- Direccion (Address): event Address

The columns to use to answer our business question are: DATE, TIME, VEHICLE_CLASS, BOROUGH, INFRACTION, VALUE

3. Methodology

3.1.Business understanding

It is important for the district to know the potential of its information, every day hundreds of subpoenas are presented throughout the city, in this case we are interested in those caused by ingestion of alcohol or other substances.

3.2.Analytic approach

A data visualization is carried out using libraries such as pandas, numpy and matplotlib, to later carry out some regressions and thus define the clusters to be evaluated geographically.

3.3.Data requirements

The data is reviewed to verify its integrity, an ETL process is carried out on the dataset, eliminating the records whose information is quite incomplete.

3.4.Data Collection

The data contains the information necessary to perform the analysis with a total of 402,500 comparisons of the year 2019, this dataset belongs to the district and its scope is public, this data set has been selected since in 2020, due to the pandemic, the data they can be very irregular.

3.5. Data Understanding

The fields that generate conclusive results are expected to be the following: DATE, TIME, VEHICLE_CLASS, BOROUGH, INFRACTION, VALUE.

3.6. Data Preparation

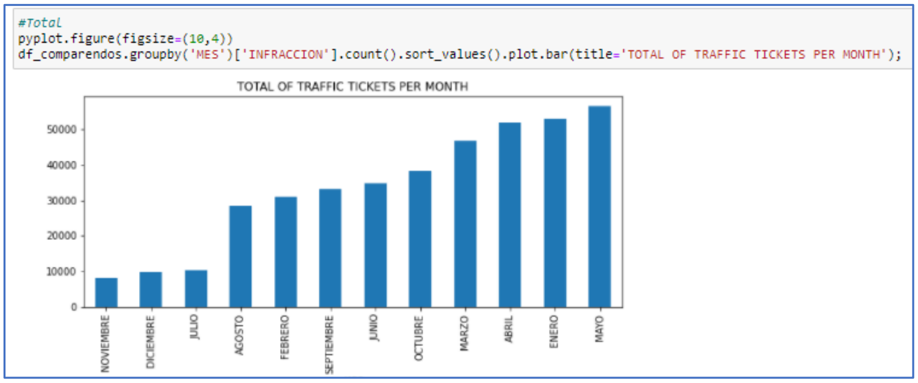
- This base is added to the value of the comparison calculated in daily minimum wages in 2019 and represented in pesos.
- The columns Year, detail comparing are withdrawn
- The latitude and longitude columns are duplicated, only one is left.
- A debugging is performed on each of the columns, converting the data without consistent information to null (N/A).
- The data correction is carried out such as unifying the localities, example MARTIRES -> Los Mártires or the vehicle class ICICLETA -> Bicycle.
- It is necessary to generate new columns from the initial data, these columns are: Grouping of the hour by intervals of 20 and 60 minutes, the type of high-level comparison and the day of the week.

4. Results

4.1.Data Modeling

Several visualizations of the information are made where the following results stand out:

- The months with the lowest number of traffic tickets are November, December and July, which are the months of festivities and a large part of the population leaves the city.



```
#Total
pyplot.figure(figsize=(10,4))
df_comparendos.groupby('MES')['INFRACCION'].count().sort_values().plot.bar(title='TOTAL OF TRAFFIC TICKETS PER MONTH');
```

Figure 1. Total of traffic tickets per month

- During off-peak hours when there is less traffic congestion, these are the hours where the most audiences are generated, both in the morning and in the afternoon

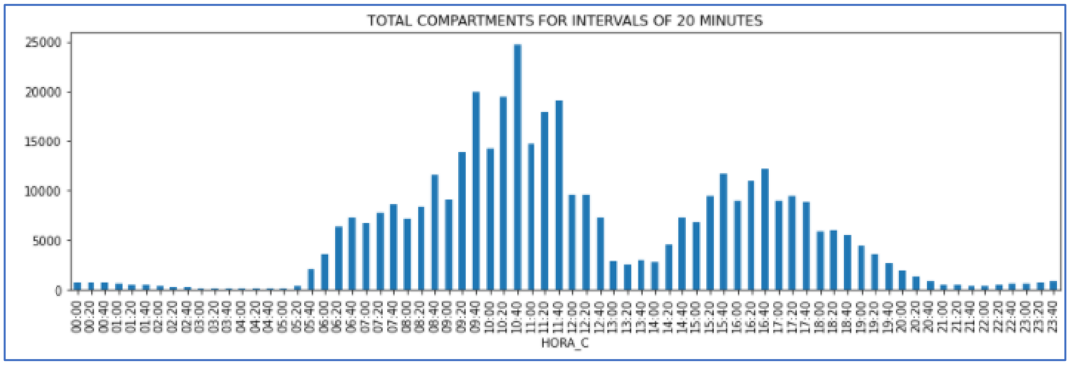


Figure 2.Total traffic tickets for intervals of twenty minutes

- For the traffic ticket type F (Driving while intoxicated or under the influence of hallucinogenic substances) the hours of the day with the highest incidence are at dawn and at night

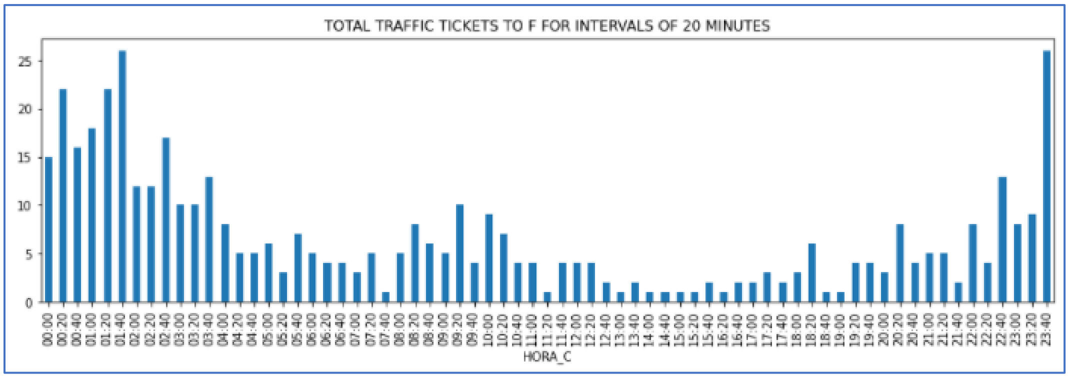
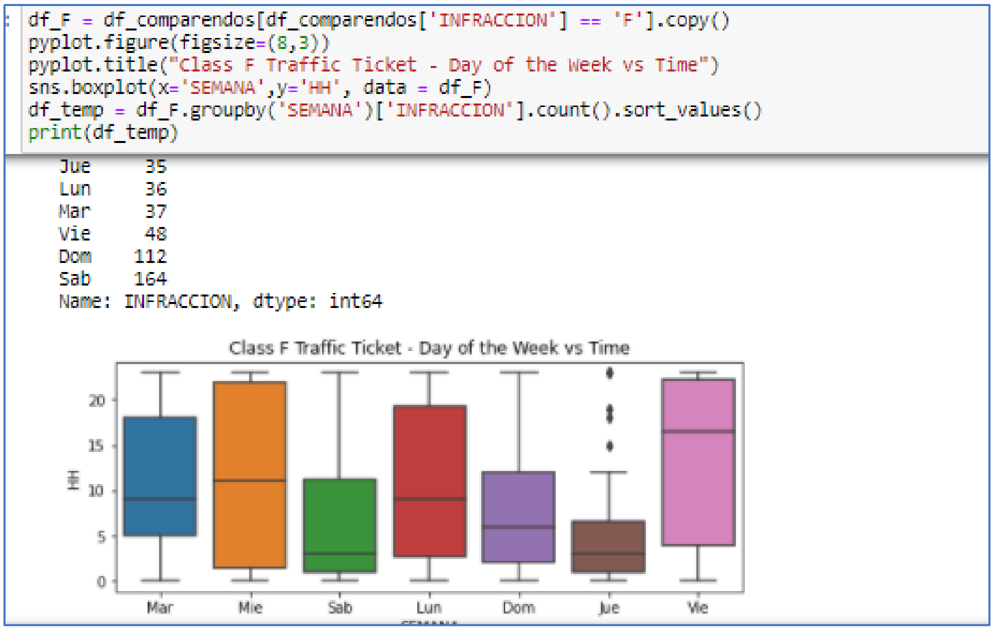


Figure 3. Total Traffic Ticket type F for intervals of twenty minutes.

- The following box diagram is divided by the days of the week and the percentiles in the hours of the F-type subpoenas. In this case, it is observed that Saturday morning is where more fines of this type are created.



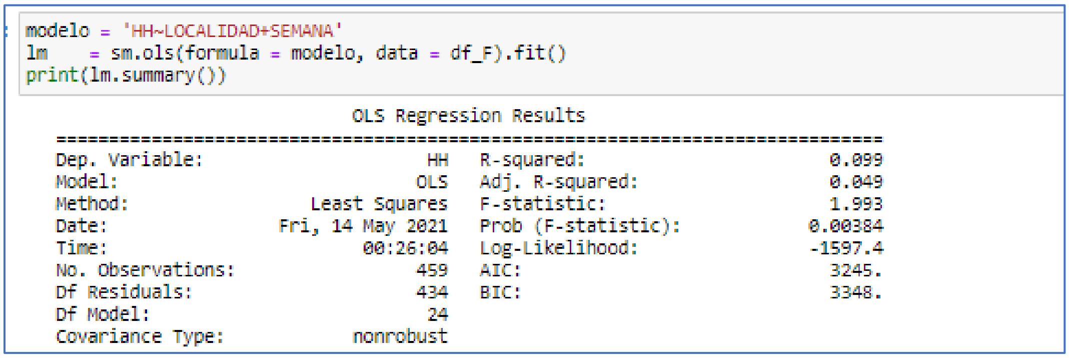
```
df_F = df_comparendos[df_comparendos['INFRACCION'] == 'F'].copy()
pyplot.figure(figsize=(8,3))
pyplot.title("Class F Traffic Ticket - Day of the Week vs Time")
sns.boxplot(x='SEMANA',y='HH', data = df_F)
df_temp = df_F.groupby('SEMANA')['INFRACCION'].count().sort_values()
print(df_temp)
```

```
Jue      35
Lun      36
Mar      37
Vie      48
Dom     112
Sab     164
Name: INFRACCION, dtype: int64
```

Figure 4. Box chart to consult the hours of the traffic ticket type F by day of the week

4.2.Evaluation

- A linear regression is carried out using the time of the offense as a dependent variable against the variables corresponding to the borough and the day of the week, it is observed that there is a coefficient of determination of 9.9% which is very low, therefore, I conclude that the generation of traffic tickets is very uniform and does not depend on the locality.



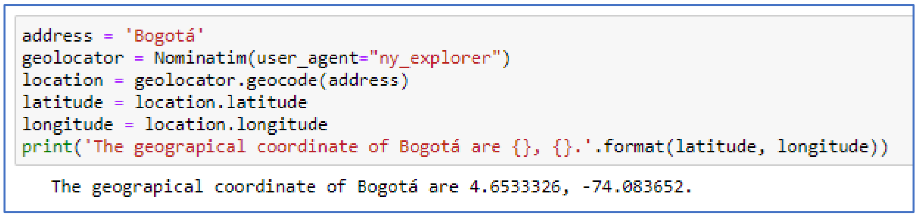
```
modelo = 'HH~LOCALIDAD+SEMANA'
lm     = sm.ols(formula = modelo, data = df_F).fit()
print(lm.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     HH   R-squared:                       0.099
Model:                            OLS   Adj. R-squared:                  0.049
Method:                 Least Squares   F-statistic:                     1.993
Date:                Fri, 14 May 2021   Prob (F-statistic):            0.00384
Time:                        00:26:04   Log-Likelihood:                -1597.4
No. Observations:                 459   AIC:                             3245.
Df Residuals:                     434   BIC:                             3348.
Df Model:                          24
Covariance Type:            nonrobust
```

Figure 5. Regression of results by borough and time of traffic tickets

4.3.Deployment

To start the exercise with foursquares I generate the geographical coordinates of the city of Bogota Colombia



```
address = 'Bogotá'
geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Bogotá are {}, {}.'.format(latitude, longitude))
```

```
   The geograpical coordinate of Bogotá are 4.6533326, -74.083652.
```

Figure 6. Coordinates of Bogotá

The following map shows the points where traffic ticket type "F" (by Driving while intoxicated or under the influence of hallucinogenic substances) were made

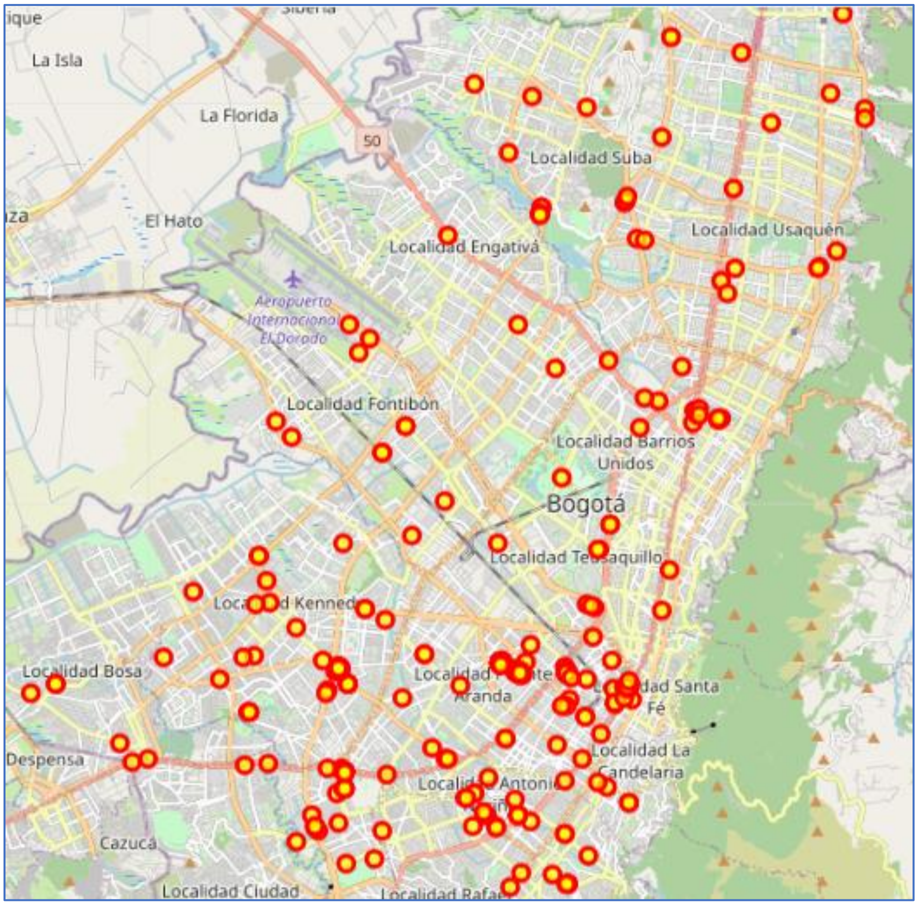
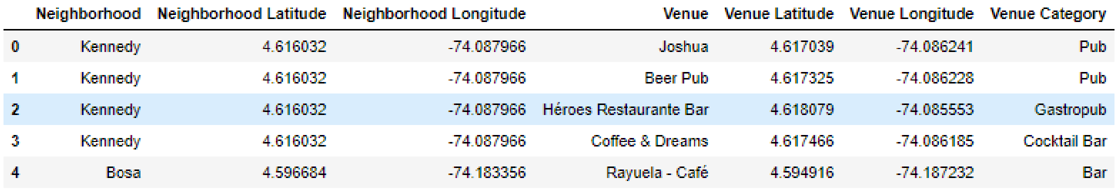


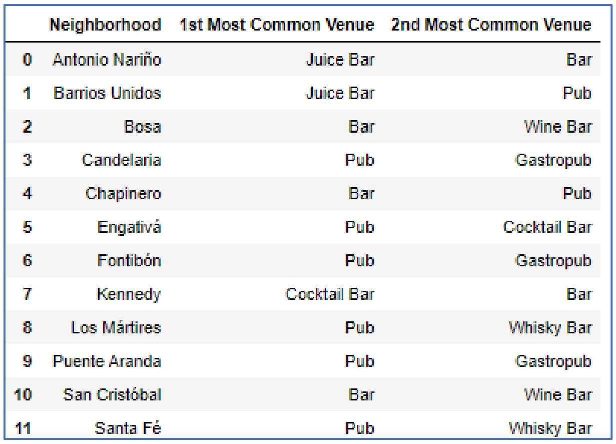
Figure 7. Coordinates of the Traffic Tickets type "F"

A search is made for bars and pubs that are close to the points where the F-type traffic tickets were generated.



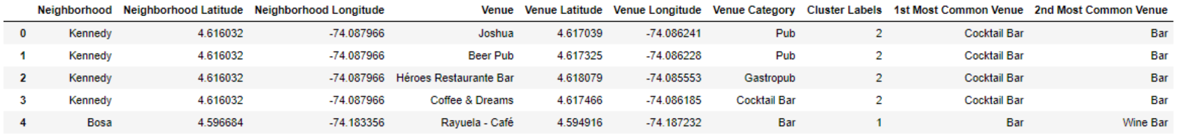
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Kennedy	4.616032	-74.087966	Joshua	4.617039	-74.086241	Pub
1	Kennedy	4.616032	-74.087966	Beer Pub	4.617325	-74.086228	Pub
2	Kennedy	4.616032	-74.087966	Héroes Restaurante Bar	4.618079	-74.085553	Gastropub
3	Kennedy	4.616032	-74.087966	Coffee & Dreams	4.617466	-74.086185	Cocktail Bar
4	Bosa	4.596684	-74.183356	Rayuela - Café	4.594916	-74.187232	Bar

For each locality, the two most common types of bars from the previous query are searched



	Neighborhood	1st Most Common Venue	2nd Most Common Venue
0	Antonio Nariño	Juice Bar	Bar
1	Barrios Unidos	Juice Bar	Pub
2	Bosa	Bar	Wine Bar
3	Candelaria	Pub	Gastropub
4	Chapinero	Bar	Pub
5	Engativá	Pub	Cocktail Bar
6	Fontibón	Pub	Gastropub
7	Kennedy	Cocktail Bar	Bar
8	Los Mártires	Pub	Whisky Bar
9	Puente Aranda	Pub	Gastropub
10	San Cristóbal	Bar	Wine Bar
11	Santa Fé	Pub	Whisky Bar

Run k-means algorithm to cluster the neighborhood into 4 clusters, Let's create a new dataframe that includes the cluster as well as the top 2 venues for each borough.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
0	Kennedy	4.616032	-74.087966	Joshua	4.617039	-74.086241	Pub	2	Cocktail Bar	Bar
1	Kennedy	4.616032	-74.087966	Beer Pub	4.617325	-74.086228	Pub	2	Cocktail Bar	Bar
2	Kennedy	4.616032	-74.087966	Héroes Restaurante Bar	4.618079	-74.085553	Gastropub	2	Cocktail Bar	Bar
3	Kennedy	4.616032	-74.087966	Coffee & Dreams	4.617466	-74.086185	Cocktail Bar	2	Cocktail Bar	Bar
4	Bosa	4.596684	-74.183356	Rayuela - Café	4.594916	-74.187232	Bar	1	Bar	Wine Bar

Finally, let's visualize the resulting clusters on the map of the city of Bogotá where it is observed how the 4 clusters are distributed by the city but in downtown a greater grouping is generated.

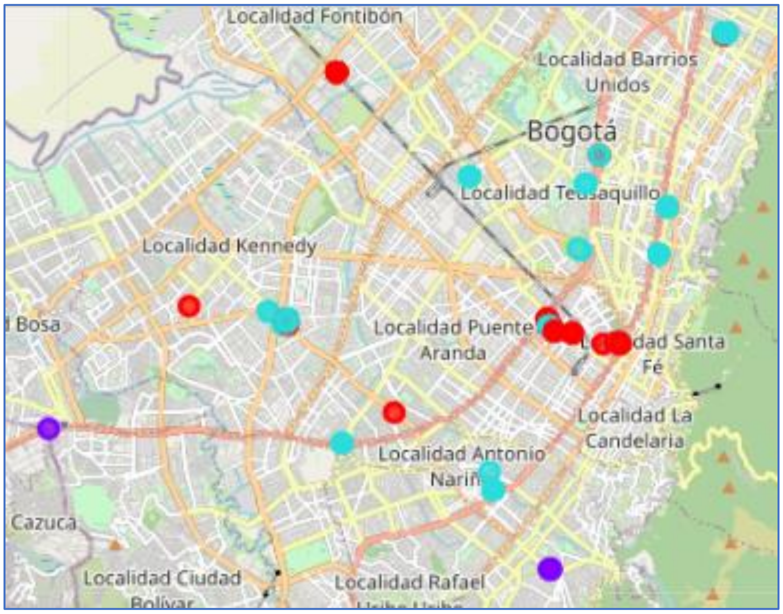


Figure 8. Four clusters created

4.4.Feedback

It is possible to generate more possibilities not only with bars and pubs but with restaurants or others, also try different values for k.

5. Discussion

The results demonstrate that around to downtown there are a higher numbers of infractions type F (Driving while intoxicated or under the influence of hallucinogenic substances). These results are interesting because it was expected a higher number around other neighborhoods like T-Zone and Gallerias.

Another explanation could be that the people who go to this type of bars have the intentions of stay there for a short period of time, chatting and eating something light, that's why they take their vehicle with them causing a higher number of incidents in that areas.

While the people who go to the other types of bars or pubs are aware that they are going to be there for longer periods, drinking and partying so they prefer do not take their own vehicles with them calling a taxi or other similar service instead.

6. Conclusion

The project demonstrates that Bars and Pubs are a sensitive point of attention, where authorities could be interested, and perhaps generate campaigns oriented to the frequent customers of these types of establishments about the consequences of driving under the effects of alcohol, preventing and reducing the numbers of infractions.

It is important to mention that the relationship between quantity of tickets and type of public establishment is not clear; other types of establishments like restaurants can have even a higher number of tickets type F around them. This first approach generates more questions for future projects where it will be possible to explore other factors and perhaps to find a clearer pattern useful to reduce the number of tickets affecting directly the cause of the problem.

7. References

[1] Julian Leonardo Martinez. github julemzc /Capstone-Toronto https://github.com/julemzc/Capstone-Toronto/blob/0c349011a967b5e1fec4daa8fa5cb1e2b530052e/Capstone%20Traffic%20Tickets%20Bogota.ipynb